Identifying a type of genetic code in an anonymous, prokaryotic DNA sequence

Aaron Pfennig[1], Alexander Lomsadze[2], Mark Borodovsky[1, 2, 3]

[1]School of Biological Sciences, [2]Wallace H. Department of Biomedical Engineering, [3]School of Computational Science and Engineering

Introduction

- It was commonly believed that genetic code was universal when it was discovered.
- Due to technological advances variations of the canonical code have been discovered.
  → Calling for an *ab-initio* approach
- Recently, phages with two different genetic codes have been described.
  → Computational tool must be able to predict potential switching points

Why is gene prediction important?

- If an outbreak of a new virus has happened:
  → accurate gene prediciton is required to help to identify potential drug targets in downstream analysis

Conclusion

- First tool of its kind
- Accurate on complete genomes and contigs greater than 10Kbp
  → makes use of other codon frequencies to determine to which amino acid a stop codon is reassigned
- Code switch point predicted with a mean error of 0.53 genes ± 6.47 genes
  → Utilizes prediction of canonical and non-canonical model to refine switching point predictions in difficult cases

Contact

- E-mail: apfennig3@gatech.edu
- Phone: +1 (404) 649 2817

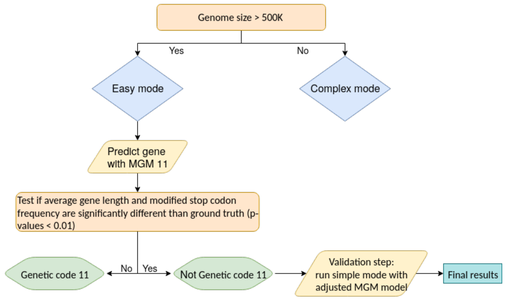
References

1. S. Osawa and T. H. Jukes, Codon reassignment (codon capture) in evolution
2. A. Rodin and S. Branciamore, The Universal Genetic Code and Non-Canonical Variants
3. N. N. Ivanova et al., Stop codon reassignments in the wild
4. N. Yutin et al., Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut
5. E. Guerin et al., Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut
6. K. Zahonova et al., An Unprecedented Non-canonical Nuclear Genetic Code with All Three Termination Codons Reassigned as Sense Codons
7. W. Zhu, A. Lomsadze, and M. Borodovsky, Ab initio gene identification in metagenomic sequences
8. A. Lomsadze, K. Gemayel, S. Tang, and M. Borodovsky, Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes.
9. L. Xu et al., Average Gene Length Is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms

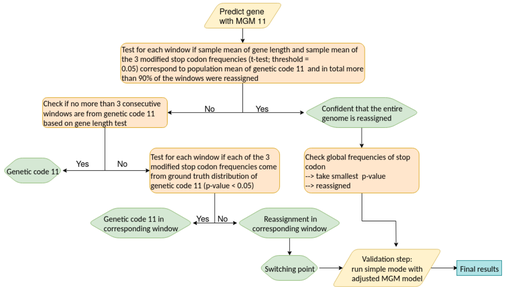
Materials and Methods

Workflow of Genetic Code Identifier

Simple mode



Complex mode



Data

- Dataset 1: 100 representative genomes with genetic code 4 and 25; GC%: 20%-40%; Size: <= 2.25Mbp
- Dataset 2: 95 representative genomes with genetic code 11; GC%: 20% - 35%; Size <= 2Mbp
- Dataset 3: 100 representative genomes with genetic code 11; GC%: 20% - 80%; Size <= 12Mbp
- Dataset 4: +5000 representative genomes used for the training of GeneMark-S2
- Dataset 5: 42 genomes of phages; GC%: 27% - 35%; Size: 175Kbp

Features

1. **Modified stop codon frequencies**

$$f(TAA)' = f(TAA) \times GC\% \quad (1a)$$
$$f(TGA)' = f(TGA) \times (1 - GC\%) \quad (1b)$$
$$f(TAG)' = f(TAG) \times (1 - GC\%) \quad (1c)$$

- If reassigned: frequency is significantly increased

2. **Average gene length**

- MetaGeneMark model with correct genetic code ~1000nt
- MetaGeneMark model with incorrect genetic code ~400nt

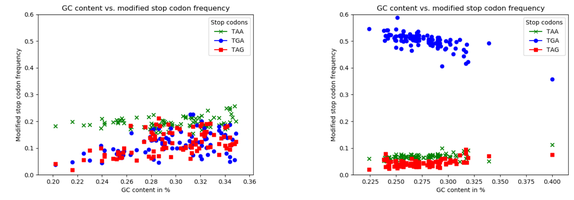


Figure 1: Modified Stop codon frequencies. TGA is reassigned in genetic code 4, its frequency is significantly increased.

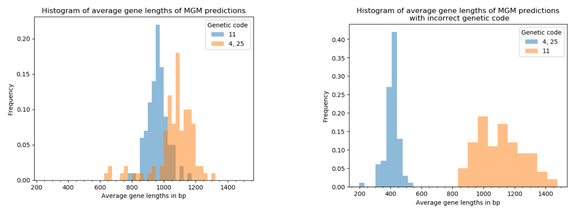


Figure 2: Av. gene lengths drops from 1000nt to 400nt if incorrect model is employed (dataset 1& 2).

Results

Simple mode	Data 1	Data 2	Data 3
No reassignment	0	95	100
Complete reassignment	100	0	0
Accuracy	1.0	1.0	1.0
Complex mode	**Data 1**	**Data 2**	**Data 3**
No reassignment	0	88	99
Partial reassignment	1	7	1
Complete reassignment	99	0	0
Accuracy	0.99	0.93	0.99

Table 1: Results on dataset 1, 2 & 3. No misclassifications were made in the simple mode. When employing the complex mode there is one genome predicted to have a partial reassignment of ~85% in dataset 1. In dataset 2 & 3 some partial reassignments of less than 10% are predicted and hence should be considered as artifacts.

1. Simple mode

- Tested on **dataset 4**:
  - 5 genomes annotated as genetic code 4 but predicted with genetic code 11 at NCBI
  - All *Acholeplasma sp.*
    - Av. Gene length ~1000nt
      → NCBI **agreed and changed** code assignment

2. Complex mode

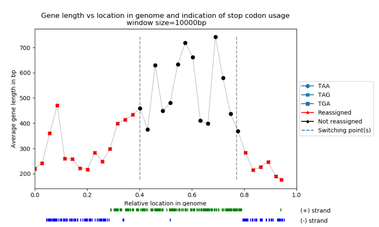


Figure 3: The analysis iof a genome of a phage in **dataset 5**. The phage is predicted to have two different genetic codes. The predictions are concordant with the literature.

- Split dataset 1 & 2 into contigs of different sizes and evaluated performance

$$\left.\begin{aligned} Specificity &= 0.99 \\ Sensitivity &= 0.92 \end{aligned}\right\} contigs\ of\ size\ 10Kbp$$

- Merged contigs from dataset 1 & 3 to simulate switching points
  - **Mean error** 0.53 genes ± 6.47 genes
  - in reality it might be less when strand information can be utilized for refinement of the prediction
    → simulated genomes do not show change in encoding as observed in the phages of dataset 5

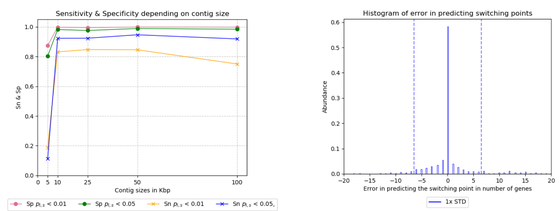


Figure 4: Evaluation of the complex mode and the switch point prediction on simulated contigs.